WGA of Obesity

Analytical Strategies and Examples

MRC Epidemiology Unit

Overview

- The initial focus will be on quality control and a summary analysis of individual SNPs, to be followed by more detailed and sophisticated ones involving hotspots and multiple SNPs and genes.
- Research papers will be written for thorough treatise on specific problems and findings. Note items under "further analysis" in the following are there only because of their labour-intensive nature but may well be topics reported in separate papers.
- I can make a note of discussion points and circulate this later

Descriptive analysis of traits

- Descriptive analysis includes summary statistics and graphics.
- A decision has to be made regarding outliers.
- The non-genetic determinants of these traits will be summarised or referred.

Single locus analysis

- Quality control including call rates
- Hardy-Weinberg equilibrium (HWE)
- Coding according to minor allele and genetic models
- Case-control comparison

Multilocus analysis on regions of interest

- Linkage disequilibrium (LD) and population characteristics
- Haplotype analysis and covariate adjustment, gene-environment interaction

Further analysis

- Genotyping error
- Sensitivity analysis
- Population substructure
- Multistage design and analysis, joint analysis
- Gene-based analysis, Hotelling's T^2 statistic and database extraction
- Analysis of pathways
- Meta-analysis with earlier reports and data from other sources.
- Retrospective versus prospective methods

Implementation

- The workhorse would be SAS/GENETICS or Stata, possibly in conjunction with customised programs in C/C++, Fortran or S-PLUS/R. This was only owing to the myriad of non-standard, individually written, statistical genetic software available. The SAS system is quite reliable, stable and available on both Windows and Linux, along with Stata and S-PLUS/R, and some programs in C/C++, Fortran. A by-product is customised software to be distributed.
- The procedures in SAS/GENETICS were mainly designed for genetic association studies of both population and family data, ranging from single-locus analysis including allele/genotype frequency calculation, Hardy-Weinberg equilibrium tests, genomic control to multilocus analysis including LD measures, haplotype estimation, as well as case-control association tests and adjustment for multiple testing. It has the added advantage of being able to utilise a variety of standard statistical procedures available. The organisation of these procedures is such that all intermediate statistics and test results can be stored as databases.

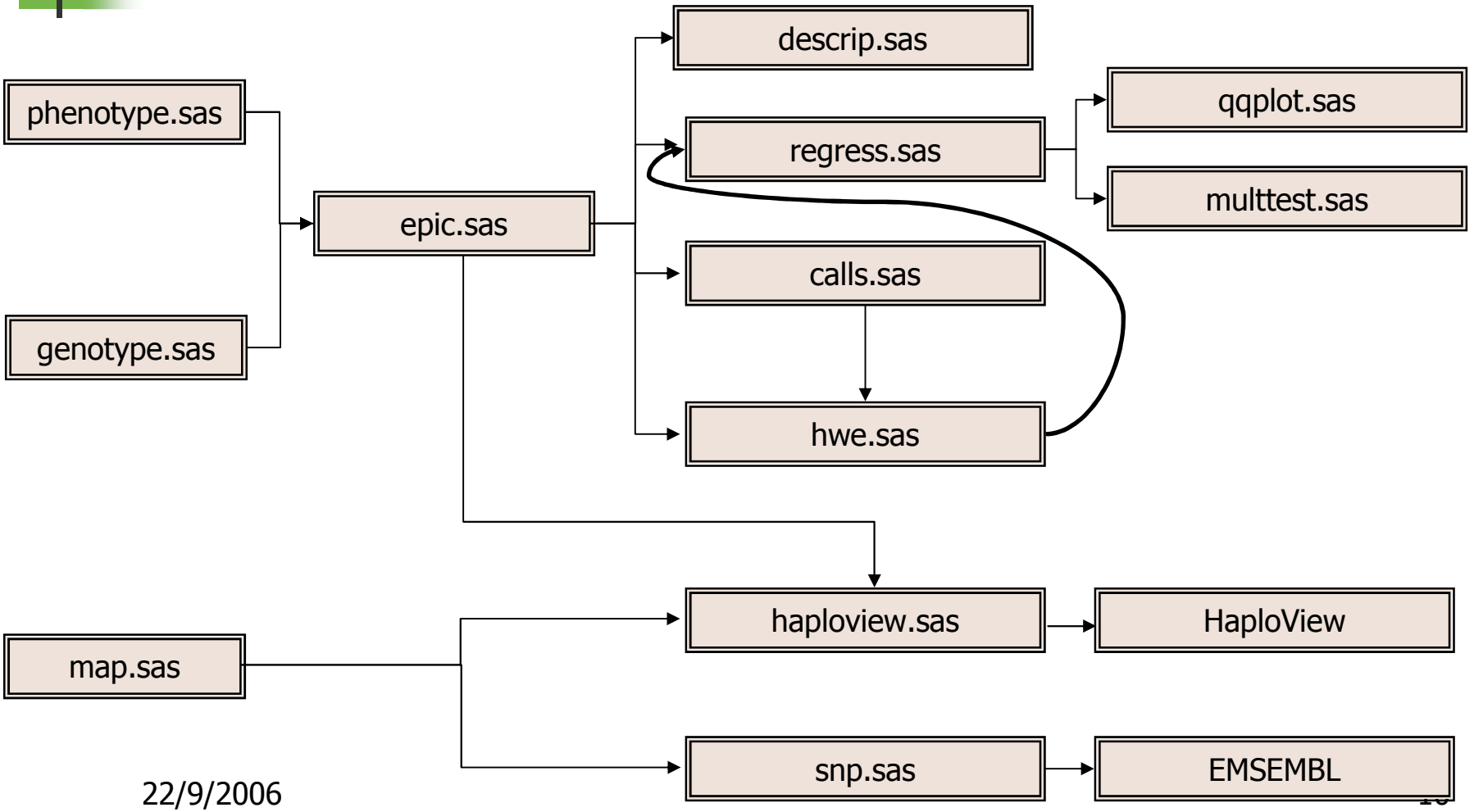
More information

- Detailed reviews of the advantage and disadvantages of genetic analysis by individual computer programs and on general statistical packages have been given (Zhao and Tan 2006, *Hum Genomics,* 2006, *Curr Bioinformatics*).
- The following section describes two examples showing prototype programs for the EPIC study of 400 controls on ~250k SNPs as well as analysis of diabetes using chromosome 20 data from Ashkenazi and four UK populations.
- One might suggest comparison of results from the SAS and Stata platforms used, but customised programming will be required for Stata.

EPIC 400 analysis

- The extensible, modular SAS programs shown in Figure 1 run without changes on both Windows and Linux systems, and are shared by all users who have access to the database.
- The results of HWE tests could be re-used for several traits including BMI, HBA1c with or without adjustment for covariates such as age, sex.
- The accommodation of both discrete and continuous traits is also straightforward.

Flowchart of the EPIC 400 Analysis



Chr 20 data

- Apart from genotype coding which was programmed in SAS, the analysis was done exclusively in Stata by Jian'an. Besides Stata's own facility, HWE tests and multiple testing can be done via publicly available routines. It has additional difficulty with covariates than framework currently available in the literature.
- Exclusion criteria include: i. SNPs with no coordinates (5) or overlap of coordinate but unable to locate from Enembl or Entrez Gene (1); ii. allele frequencies < 1% in combined data (270); iii. HWE in cases <0.00001 or controls <0.0001 in combined data (190); iv. Call rate < 80% in individual cohort or <90% in combined data (86); v. p<0.0001 between stages 1 and 2 allele frequencies in either cases or controls in individual cohort or combined (37). A total of 589 SNPs were removed from the dataset (out of 4546=4544 at phase 1 + 751 at phase 2).

Hotelling's T^2 statistic

- It can be derived as a score statistic, discussed in the context of staged design, and can be used to test for multiple SNPs while accounting for correlations them.
- In the case of linkage equilibrium the score statistic involving several neighbouring loci reduces to the sum of individual chi-squares.
- In fact the sum statistics have been exclusively studied by Ott and associates.

Joint analysis of data from two and three stages

- An earlier work was due to Lowe et al. (2004) Gene Immun which included a multistage framework called stopping for futility. The paper nevertheless focused on the design.
- A following paper by Skol et al. (2006) Nat Genet showed joint analysis is more powerful than replication study. It uses test statistics involving several disease models and does not take into account correlations between SNPs.
- A recent approach was described by Lin (2006) Am J Hum Hered, which also accommodates dominant/recessive/additive recoding on genotypes with significant levels obtained from Monte Carlo simulation. Other work by Lin and colleagues was haplotype analysis of a variety of study designs including case-cohort.

SAS/GENETICS

- The module consists of procedures ALLELE, CASECONTROL, FAMILY, HAPLOTYPE, HTSNP, INBREED, PSMOOTH, TPLOT which implement allele, genotype and haplotype frequency estimation and tests for differences appropriate for unrelated individuals and family data. In conjunction with other procedures such as LOGISTIC, GLM, GENMOD, MIXED, PHREG it provides comprehensive and integrated environment for analysis. The database, graphics, programming, Internet facility, among others, is well-documented.

Summary Statistics

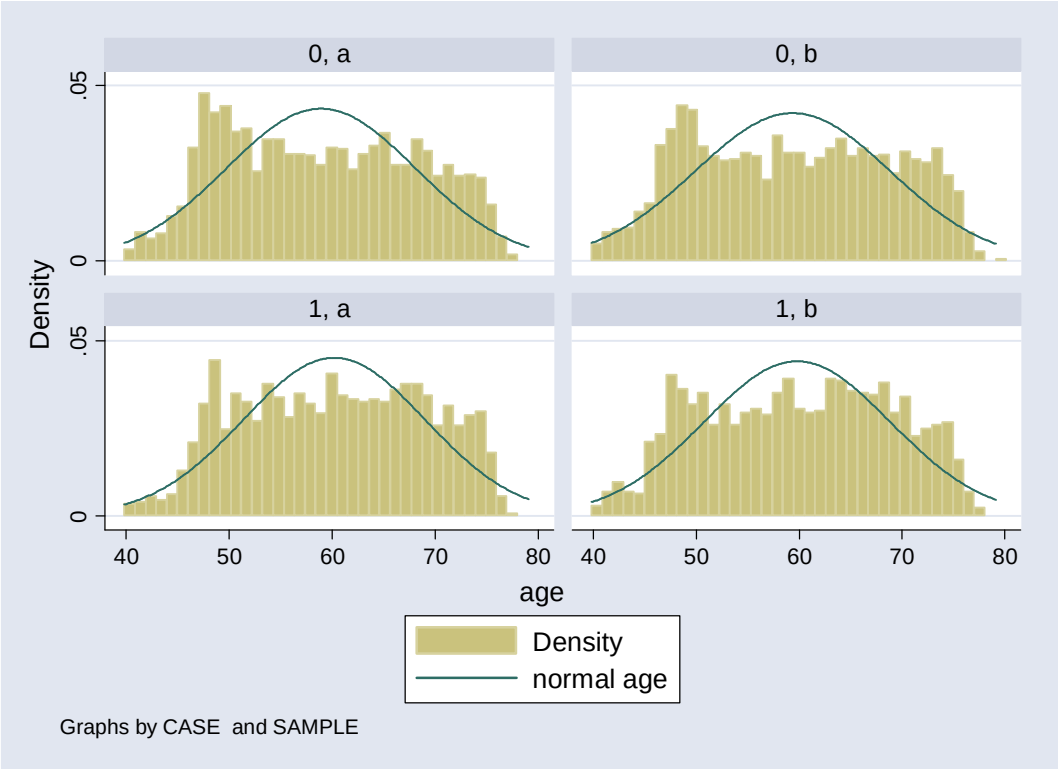
- **. count**
- 7686

- **. table cohort case sample, col row**

	SAMPLE and CASE					
	-------- a --------			-------- b --------		
COHORT	0	1	Total	0	1	Total
0		1,343	1,343		1,343	1,343
1	2,124	376	2,500	2,137	363	2,500
Total	2,124	1,719	**3,843**	2,137	1,706	**3,843**

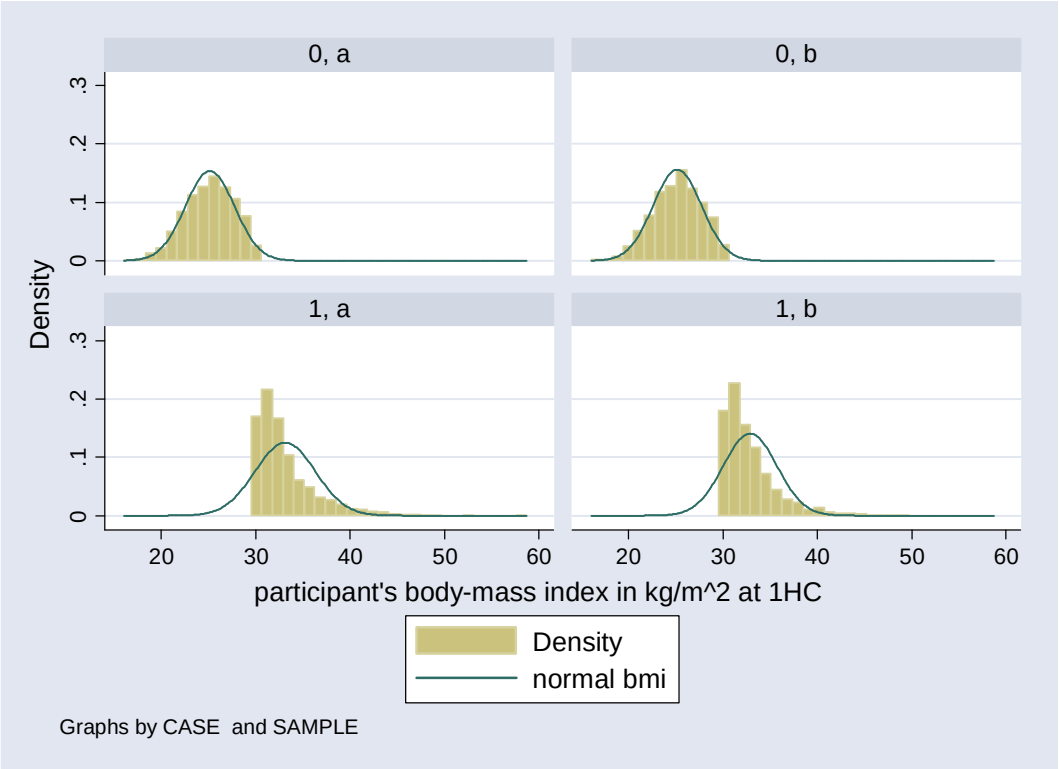
The age distribution

. histogram age, normal by(case sample)



The distribution of BMI

. histogram bmi, normal by(case sample)



Further summary

. sum bmi

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         bmi |      7681    28.64962     4.79068   16.06659   58.69388
```

. table obesity obesity2

```
           |   obesity2
   obesity |     0      1
-----------+-------------
         0 | 2,508    169
         1 |   222  1,637
```

. table cohort obesity2 obesity, col row

```
           |            obesity and obesity2
           | -------- 0 --------    -------- 1 --------
    COHORT |     0      1  Total        0      1  Total
-----------+----------------------------------------------
         0 |                           175  1,287  1,462
         1 | 2,508    169  2,677        47    350    397
           |
     Total | 2,508    169  2,677       222  1,637  1,859
```

Process Steps for WGA Obesity Analysis

SETTINGS OR PROCESS STEPS	Phase 1 QC	Phase 2 Single-point	Phase 3 Multipoint	Phase 4 Other	Phase 5 Report
DETAILS	QC HWE	Binary trait	Haplotype analysis	Genotyping error Sensitivity analysis Population structure Joint analysis Pathway analysis Meta-analysis Comparisons	Result Statistics
OUTCOMES/GOALS	**High quality data**	**Standard**	**Further information**	**Further information**	**Paper**
TIME LINE	October?	October?	November?	December?	2007?